

REMARKS

This is in response to the Office Action mailed on November 13, 2006. In the Office Action, claims 1-8 and 25-34 were pending.

Claim 2 was objected to because the status identifier was incorrect and included a typographical error. With this amendment, the status identifier typographical error have been corrected. Thus, withdrawal of the objection to claim 2 is respectfully requested.

Claims 1-8 and 25-34 were rejected under 35 U.S.C. § 112, first paragraph, as failing to comply with the written description requirement. In particular, the Office Action reports that claim 1 recites the limitations of “extracting a first set of related elements” and “extracting a second set of related elements”, wherein said limitations have not been found in the specification and thus constitute new matter. Applicants respectfully disagree. On page 11 of the specification, at line 17, elements are included in extraction patterns and, on line 22, are defined as “variables containing information related to a particular subject...”. Furthermore, on page 12, the specification describes, “information extraction is concerned with extracting information related to a particular subject. Extracted information can include pairs, triplets, etc. of related elements pertaining to the subject.”

The pairs, triplets, etc. of related elements can be referred to as a set. Applicant respectfully submits that a “set” is commonly referred to as a group of things of the same kind that belong together and are so used. Thus further, on page 12, several pairs (constituting a set of related elements) of information that can be extracted include title and author information, inventor/invention information, question/answer pairs, etc. As one skilled in the art can readily attest, multiple sets of related elements can be extracted from an information source using, for example, extraction module 200 of FIG. 2 and FIG. 3. Applicant has further amended “topic” to “subject” to more closely correspond to the specification. Based on the foregoing, it is believed that both extracting a first set of related elements and extracting a second set of related elements are adequately described such that one skilled in the art can reasonably understand the claimed invention. Thus, withdrawal of the rejections under 35 U.S.C. § 112 is requested.

Claims 1-8 and 25-34 were rejected under 35 U.S.C. § 103(a) as being unpatentable over Yangarber et al. (2000) in view of Soderland (1999). Yangarber et al. describes an information extraction system that utilizes pattern matching to extract a set of pairs

(see section 3.3) having a verb positioned therebetween (see section 3.4). The patterns are “generalized” in that they refer to a semantic class, which can include a number of words. Documents that have words that match the patterns of semantic classes can have information contained therein be extracted. However, Yangarber et al. does not teach or suggest including both words and wildcards positioned between semantic classes.

Soderland describes learning information extraction rules and discusses a wildcard, which “means to skip any number of characters until the next occurrence of the following term in the pattern”. A restricted wildcard is also discussed that only skips characters in a semantic field. Soderland does not teach or suggest restricting the wildcard to less than a specified number of words.

In contrast, subject matter disclosed in the present application is directed to information extraction from extraction patterns of related elements, words and wildcards. The words and wildcards are positioned between the related elements and wildcards denote less than a specified number of words to be skipped. As discussed in the application, extracting information from a source is performed to output related elements pertaining to a subject from the patterns. For example, a company/product pair can be extracted from documents that are related to a product release. By limiting a number of words that can be skipped, an improved extraction module can be realized.

In view of these differences, applicants have amended independent claims 1 and 5 to clarify the features recited therein. Claim 1 has been amended to recite a computer-implemented method of extracting information from an information source comprising a plurality of documents. The method includes accessing strings of text in the information source and comparing the strings of text in the information source with generalized extraction patterns. A plurality of strings in the information source are identified that match at least one generalized extraction pattern. The generalized extraction patterns include related elements pertaining to a subject, words and wildcards, wherein the wildcards denote that at least one word and less than a specified number of words in an individual string can be skipped in order to match the individual string to an individual generalized extraction pattern. The words and wildcards are positioned between the related elements. The method also includes extracting a first set of related elements of text pertaining to the subject from a first string of the plurality of strings based the related

elements pertaining to the subject in the at least one generalized extraction pattern. The first string is associated with a first document in the plurality of documents. The method also includes extracting a second set of related elements of text pertaining to the subject from a second string of the plurality of strings based on the related elements in the at least one generalized extraction pattern. The second string is associated with a second document in the plurality of documents. At least one of the related elements of text in the first set of related elements is different from each of the related elements of text in the second set of related elements of text. The first related set of elements and the second set of related elements are output.

Similarly, independent claim 5 has been amended to recite a computer-readable medium for extracting information from an information source comprising a plurality of documents. The medium includes a data structure including a set of generalized extraction patterns including related elements pertaining to a subject, words and an indication of a position for at least one optional word and less than a specified number of words. The words and the at least one optional word are positioned between the related elements. An extraction module uses the set of generalized extraction patterns to match a first string and a second string in the information source with one of the generalized extraction patterns. The first string is associated with a first document in the plurality of documents and the second string is associated with a second document in the plurality of documents. The extraction module also extracts a first set of related elements of text pertaining to the subject from the first string based on related elements in said one of the generalized extraction patterns and a second set of related elements of text pertaining to the subject from the second string based on the related elements in said one of the generalized extraction patterns. At least one of the related elements of text in the first set of related elements is different from each of the related elements of text in the second set of related elements of text. The extraction module also outputs the first related set of elements and the second related set of elements.

Features recited in claims 1 and 5 are neither taught or suggested by the combination of Yangarber et al. and Soderland. In particular, the features in the claims relate to extracting information using extraction patterns of related elements, words and wildcards that indicate words can be skipped and denote less than a specified number of words that can be skipped.

In view of the foregoing, Applicants submit that the present application is in condition for allowance. Withdrawal of the rejections and allowance of the pending claims is respectfully requested.

The Director is authorized to charge any fee deficiency required by this paper or credit any overpayment to Deposit Account No. 23-1123.

Respectfully submitted,

WESTMAN, CHAMPLIN & KELLY, P.A.

By: 

Todd R. Fronek, Reg. No. 48,516  
900 Second Avenue South, Suite 1400  
Minneapolis, Minnesota 55402-3319  
Phone: (612) 334-3222 Fax: (612) 334-3312

TRF:rdg